Codificación no paramétrica de voz basada en redes neuronales sobredimensionadas

Gustavo Hernández-Ábrego, Eloi Batlle, Carles Antón, Enric Monte[1]
DEPARTARTAMENTO DE TEORÍA DE LA SEÑAL Y COMUNICACIONES
UNIVERSITAT POLITÈCNICA DE CATALUNYA
c/GRAN CAPITÀ, s/n, 08034, BARCELONA
Correo electrónico: {abrego/eloi/carles/enric}@gps.tsc.upc.es

Abstract:

This paper presents a non-parametric voice coding system built with a Multi-layer Perceptron (MLP) whose hidden layer cointains less neurons than the outer ones. In the hidden layer, a compressed representation of the voice signal is obtained. This system has been designed in a way that the neural network that compresses the signal is also useful to uncompress it. The computational complexity of a network such as this is very high and the compression-expansion process decreases the signal quality. In order to cope with these two drawbacks, the MLP is over dimensionated adding signal related data as hints for the training process. A number of hints are tried and experimental tests show that overall performance increases with the add of hints.

1. Introducción

Un sistema de compresión de datos se puede implementar en base a un Perceptrón multi capa (MLP) de acuerdo a la arquitectura propuesta por Cottrell/Munro/Zipser [1]. Esta arquitectura consiste en establecer capas de entrada y salida del mismo tamaño y una capa intermedia de tamaño inferior. Esta técnica de compresión está relacionada con una transformada en la cual los pesos de las neuronas que conectan a la capa de entrada con la capa intermedia pueden considerarse como la matriz de transformación y los pesos de las neuronas que conectan a la capa intermedia con la capa de salida como la transformada inversa, tal como se muestra en la Fig. 1. De tal forma, la representación interna (la activación de las neuronas en la capa intermedia) es una versión comprimida de la señal de la entrada. Los pesos de una red así se pueden calcular en base a un entrenamiento del tipo *back propagation* [2]. Este tipo de entrenamiento está basado en una búsqueda de gradiente, tendiente a caer en mínimos locales que repercuten en problemas de convergencia en la red y en la disminución de la calidad de la señal. Con el fin de evitar estos problemas, se ha empleado la técnica propuesta en [3] que consiste en emplear *hints* (pistas) referentes a la función aprendida en la fase de entrenamiento. Estos hints son empleados como si fueran datos de la función de aprendizaje, de manera que la red tenga que minimizar el error entre esta función y la salida deseada y entre los hints y la salida deseada al mismo tiempo. Esta es una manera de sobredimensionar el problema del entrenamiento, dando más información referente a la señal que la que la señal misma ofrece. Una vez que la red ha sido entrenada, los pesos de las unidades de salida relacionadas con los hints no son usados, por lo cual la carga computacional del sistema no se incrementa. Este sistema ya ha sido probado en [4] con resultados que prueban que el uso de hints favorece la convergencia del entrenamiento y aumenta la calidad de la señal obtenida en un sistema de compresión y expansión. En este trabajo se hace énfasis en la información de alta frecuencia contenida en la señal utilizando para ello hints que resalten dicha información con objeto de obtener una mejor representación de la señal de voz.

La siguiente sección describe la configuración del sistema, en lo tocante a la base de datos empleada y al tipo de experimentos realizados. A continuación se describe la topología de la red neuronal empleada en este trabajo. Los diferentes tipos de hints empleados se consideran en la sección 4. La sección 5 muestra los datos experimentales desprendidos de las pruebas realizadas con el sistema. La sección 6 presenta nuestras conclusiones.

2. Descripción general del sistema

El sistema se forma a partir de un MLP formado por 3 capas: una capa de neuronas de entrada, una de neuronas de salida y una capa de neuronas ocultas. Para que este esquema funcione como un sistema de codificación , la capa oculta debe de tener un número menor de neuronas que las capas externas. El cociente entre el número de neuronas a la entrada y el número de neuronas a la salida será la razón de compresión. Ya en un estudio previo [4], se había observado que el mejor sitio, en términos de la calidad de la señal tratada, donde agregar los hints es la capa de salida del MLP. Este sistema se muestra de manera esquemática en la Fig 1. Como preprocesado de la señal, se utiliza un filtro de respuesta impulsional finita (FIR) cuya función de transferencia se expresa en (1).

$$H_p(z) = -0.04 + 0.31z^{-1} - 1.06z^{-2} + 2.5z^{-3} - 1.06z^{-4} + 0.31z^{-5} - 0.04z^{-6} \quad (1)$$

Este filtro enfatiza el rango de alta frecuencia pero sin atenuar el de baja frecuencia que

[1] Este trabajo ha sido financiado en parte por el Conacyt y en parte por el gobierno español con el número de convenio TIC95-1022-C05-03.

es donde se concentra el 70 % de la energía de la señal. Este es un hecho considerable ya que el algorítmo de back propagation basa su convergencia en los niveles de energía que puede aprender de los ejemplos. El filtro que devuelve las características espectrales a la señal tiene una función de transferencia que se muestra en (2). Es una aproximación FIR al filtro inverso de (1).

$$H_d(z) = 0.002 - 0.03z^{-1} + 0.14z^{-2} + 0.7z^{-3} + 0.14z^{-4} - 0.03z^{-5} + 0.002z^{-6} \quad (2)$$

La señal enfatizada, será la entrada al MLP. En la etapa de entrenamiento, la salida también será la señal preenfatizada pero además se agregaran los hints. En el entrenamiento, los pesos de las tres capas de la red se ajustan en base a una función de coste multidimensional por medio de una estrategia de back propagation. El error entre la señal de salida deseada y la que se obtiene cada vez de la red, será el indicador de la evolución del entrenamiento. Cuando se alcance el nivel de error deseado. Se detiene el entrenamiento y se fijan los valores de los pesos de cada neurona de la red. Esta red deberá de funcionar al mismo tiempo como sistema compresor y descompresor. La etapa de compresión la forman la capa de entrada, y la capa intermedia. La salida de esta etapa es la señal comprimida. La etapa de descompresión la forman la misma capa intermedia y la capa de salida. La entrada a esta etapa será la señal comprimida y la salida será la señal reconstruida más la salida de las neuronas que consideraban a los hints, cuya salida es poco importante y sólo sirve para proporcionar información a la reconstrucción.

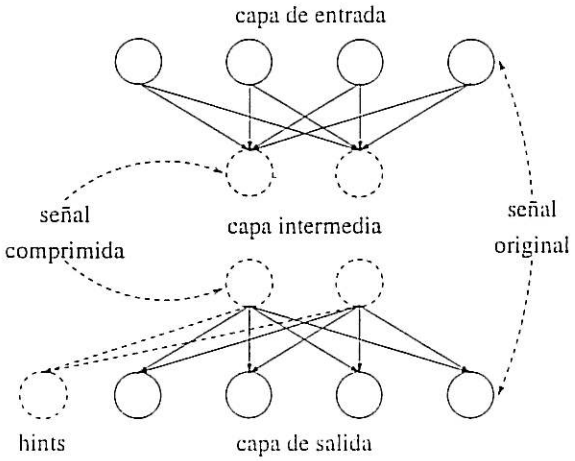

Fig. 1 MLP con hints como sistema de codificación

2.1 La base de datos

Para las fases de entrenamiento y prueba del sistema, se ha utilizado parte de la base de datos TIMIT, que contiene frases fonéticamente balanceadas pronunciadas en inglés y grabadas por locutores adultos de ambos sexos. Esta base está muestreada a 16 kHz. Para el entrenamiento, se han empleado dos de estas frases provenientes de distintos locutores. Se ha utilizado otra de las frases para la validación y evitar el sobre-entrenamiento. Para la fase de prueba del sistema, se han empleado dos de las frases de TIMIT que nada tienen que ver con las frases empleadas en el entrenamiento y en la validación.

La señal de voz, se divide en tramas de longitud variable. Estas tramas servirán de ejemplos de entrada a la red para el entrenamiento.

2.2 Diseño de los experimentos

Los experimentos se diseñan para probar cuales son las prestaciones que cada tipo de hints puede aportar al sistema. Para cada prueba, se calculan los hints referentes a la señal de entrenamiento, se agregan a la capa de salida del MLP y se realiza el entrenamiento de esta red. Cada vez que se quiera cambiar el tipo de hint, se ha de hacer un nuevo entrenamiento de la red.

3. Diseño de la red

La red neuronal que se emplea en este trabajo es un perceptrón con una capa oculta. Para llevar a cabo el entrenamiento de esta red, los parámetros que la describen se ajustan en base a el algorítmo de back propagation que ya ha demostrado sus prestaciones en trabajos similares [1,4]. Para acelerar el proceso de entrenamiento, se agrega en el back propagation un factor de aprendizaje (*learning rate*) adaptable, con valor inicial de 0.1, y un factor fijo de moméntum de 0.795 para la búsqueda del mínimo en la función de coste. Los parámetros (pesos y valores de sesgo) de cada una de las neurona de la red se inicializan de manera aleatoria.

La topología del MLP se hace teniendo en cuenta que cada una de las neuronas de la red, considera a un valor de la representación temporal de la señal. Por esto, el número de neuronas en la capa de entrada depende de la extensión, en términos de muestras de señal, de las tramas que se emplean. En [4] se prueban varias longitudes para la trama de señal. Los resultados muestran que la elección de tramas de 128 muestras es la que mejores prestaciones brinda al sistema. La longitud de la capa intermedia está dada en base a la razón de compresión del sistema. Por ejemplo, cuando se tienen 128 neuronas a la entrada y se desea una razón de compresión de 4, la capa intermedia tendrá 32 neuronas. El número de neuronas en la capa de salida será igual al de la entrada más el número de hints que se agreguen.

Dado el elevado número de parámetros a entrenar, y la cantidad de conexiones que puede haber entre las neuronas que forman la red, se han implementado algunas modificaciones en el algorítmo back

propagation con la intención de acelerar la fase de entrenamiento, y son los siguientes:

• División del factor de aprendizaje por el número total de parámetros que forman a la capa previa. Esta operación se realiza con el fin de hacer que la magnitud del gradiente sea la misma en todas las capas de la red. Este procedimiento es necesario en este sistema debido a que las capas externas tienen un número de neuronas varias veces mayor que el de la capa oculta, por lo que el gradiente de la capa oculta será varias veces menor al de las capas externas; con esto se normaliza la magnitud del gradiente en las distintas capas. Si no se llevara a cabo esta normalización, la razón de aprendizaje de la red entera sería mucho menor y la velocidad de convergencia sería significativamente menor.

• División del vector de valores de sesgo. Cada uno de estos valores es dividido por un factor de 4 para asegurar que el gradiente en cada unidad sea uniforme. En las ecuaciones del algorítmo back propagation, el término correspondiente a los pesos que no tienen sesgo son multiplicados por una expresión que siempre es menor que 1 (cuando la función de no linealidad de cada neurona satura a 1). Al hacer esta división, se igualan las condiciones entre los pesos que son sesgados y los que no lo son. Esta modificación empírica, ha demostrado una mejora importante en la convergencia del algorítmo.

4. Hints

Ya en [4] se apuntaba la necesidad de seleccionar con particular cuidado la información adicional que se incluya en la red como hints. No cualquier información relacionada con la señal podrá emplearse como hint. Como ya se ha mencionado antes, la energía en la señal es de particular importancia en un entrenamiento de back propagation. Por ello, hemos decidido emplear como información adicional, que se agregará en la capa de salida de la red neuronal, información referente a la energía de cada trama de la señal de voz. La energía de cada trama, se ha obtenido en base a:

$$E = \frac{\sum_{i=1}^{n} x^2}{n} \qquad n = \text{número de muestras} \quad (3).$$

Cuando n es igual al número de muestras que contiene la trama, se obtiene un promedio de la energía de la trama. Si n es menor que el número de muestras de la trama, se obtiene un vector de niveles de energía que expresan con mayor precisión la evolución de este dato dentro de la trama, es lo que llamamos "energía corta". Para las pruebas experimentales, la energía corta es un vector de 10 niveles de energía para cada trama. El dato de la energía es dinámico con respecto al tiempo. Para considerar la variación de este dato a lo largo del tiempo, se han aproximado la 1ª y 2ª derivadas de la energía obteniendo la diferencia entre los niveles de energía de tramas consecutivas.

5. Resultados experimentales

La red neuronal es entrenada y sus parámetros se fijan. Para probar sus prestaciones, se emplean señales de voz incluidas dentro de la base de datos TIMIT que no han sido presentadas a la red en la fase de entrenamiento, se hace pasar a la señal por toda la red neuronal, simulando el proceso completo de codificación y decodificación, la señal obtenida al final de la red se compara con la señal original y se evalúan las diferencias. Las diferencias se expresan en términos cuantitativos, error cuadrático medio entre las señales, y cualitativos, criterio perceptual de la señal. Con el fin de tener una referencia para evaluar las prestaciones del sistema, se comparan los resultados obtenidos con los que se desprenden de una topología de red neuronal similar pero sin información adicional agregada. La siguiente tabla muestra los resultados obtenidos por el sistema cuando se establece una razón de compresión de 1:4; tramas de 128 muestras y se emplean como hints los datos de energía. Con el nombre "energía", se hace alusión a un vector formado por el promedio de energía de la trama completo y por sus 1ª y 2ª derivadas. "Energía corta" se refiere a un vector de 30 elementos que contiene 10 valores de energía para cada trama más las aproximaciones a las 1ª y 2ª derivadas.

Prestaciones del sistema		
hint	error	calidad
ninguno	0.5596	buena
energía	0.4916	buena
energía corta	0.4358	muy buena

Los resultados experimentales muestran que los mejores resultados se obtienen cuando se agrega la energía corta en la etapa de salida de la red.

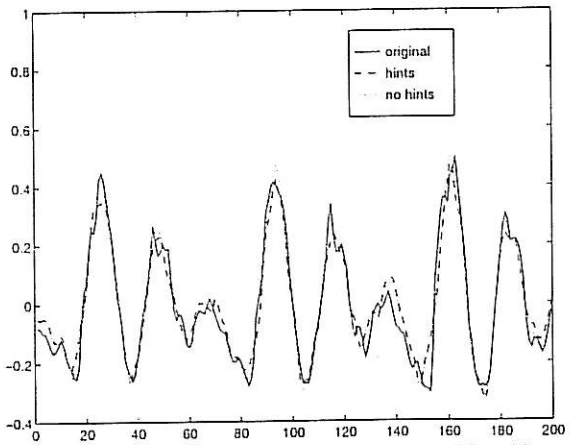

Fig. 2. Señales obtenidas del proceso de codificación

En Fig. 2, se puede observar la diferencia entre las señales que se obtienen de este sistema

cuando se agregan hints y cuando no, comparado con la señal original. Con esta representación, es difícil decir cual de las señales obtenidas es mejor, pues las dos presentan comportamiento similar. Para tener otro criterio de evaluación, se estima el espectro las señales (por medio del cálculo de 10 coeficientes LPC) obtenidas por el sistema y se comparan con la estimación del espectro de la señal original, dicha comparación se muestra en la Fig. 3.

En [4] ya se apuntaba que uno de los mayores problemas que presenta este sistema es el manejo de las altas frecuencias. El pre-énfasis que se ha agregado como preprocesamiento, está pensado para manejar esta situación. Aún así, cuando no se hace uso de hints, el sistema pierde gran cantidad de información en la banda de las altas frecuencias, como se puede observar en la Fig. 3. Las gráficas de los espectros, muestran que el sistema tiene un buen desempeño en la banda de las bajas frecuencias, inclusive sin el empleo de hints. Esto se explica con el hecho de que el entrenamiento se ha realizado en base a la energía de la señal. Cuando se emplea la energía corta en el entrenamiento del sistema, el contenido frecuencial en la región superior del espectro, se conserva mejor.

Con respecto a el criterio de calidad perceptual, es fácil distinguir entre la señal original y las señales obtenidas del sistema. La señal de referencia, presenta un mayor nivel de ruido donde tendrían que existir silencios. La señal con hints presenta un contenido frecuencial mas alto, pero no tan completo como el que presenta la señal original. Cabe destacar, que en ninguno de los casos, el sonido es "mecánico" (como podría ser el sonido cuando se trata de un esquema de codificación paramétrico que emplee sintetización) y que la forma de onda de la señal no presenta discontinuidades que pudieran resultar en ruidos y chasquidos. En cualquiera de los casos, el significado de la frase codificada fue perfectamente comprensible.

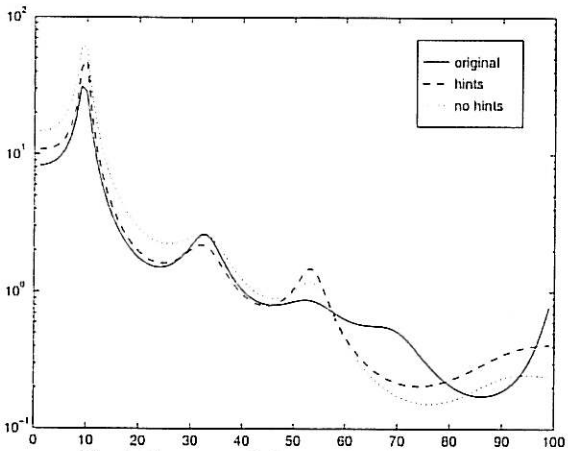

Fig. 3. Espectros de las señales del sistema.

6. Conclusiones

En este trabajo, se emplea una red neuronal para llevar a cabo una tarea de compresión de datos, entendida como la codificación de una señal de voz en una representación reducida. Nuestro esquema de codificación es no paramétrico y utiliza la arquitectura de un MLP con una capa oculta. El entrenamiento de una red neuronal así, plantea varios problemas, sobre todo en lo que a convergencia y calidad de los mínimos de la función de coste se refiere. Para reducir dichos problemas, proponemos la inclusión de información adicional (hints) a la fase de entrenamiento del sistema, para sobredimensionar la función de coste y conseguir un mejor desempeño del sistema. En base a la experimentación, se puede concluir que los hints que en mayor grado benefician a el comportamiento del sistema son los que se refieren a los niveles de energía de la señal. La señal codificada y reconstruida con este sistema, presenta alta similitud a la señal original, con un contenido frecuencial más rico que el que se puede obtener con un sistema similar pero sin la inclusión de hints. Con este trabajo, se demuestra la viabilidad de un MLP como la base de un sistema de codificación de señales de voz.

Referencias

[1] Cottrell, G. W., Munro P., y Zipser, D. "Image compression by back propagation: An example of extensional programming". *ICS Report 8702*, University of California (1987).

[2] Kung, S. Y. *Digital Neural Networks*. Prentice Hall (1993).

[3] Suddarth, S. C., y Kergosien, Y. L. "Rule-injection hints as a mean of improving network performance and learning time", *Lecture notes in computer science. Proceedings of the EURASIP Workshop Engineering* Springer Verlang (1990).

[4] Hernández-Ábrego, G., Monte, E., y Mariño, J. B., "Non parametric coding of speech by means of a MLP with hints", *Proccedings of IWANN'97* Lanzarote (1997)

[5] Furui, S. *Digital speech processing, synthesis and recognition*, Marcel Dekker (1989).